DOCUMENT RESUME

ED 397 635 FL 023 728

AUTHOR Strong, Gregory
TITLE A Survey of Issues and Item Writing in Language Testing.
PUB DATE Dec 95
NOTE 34p.
PUB TYPE Reports - Descriptive (141) -- Journal Articles (080)
JOURNAL CIT Thought Currents in English Literature; v68 p281-312 Dec 1995

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cloze Procedure; Difficulty Level; *English (Second Language); Foreign Countries; Interrater Reliability; *Language Tests; Multiple Choice Tests; Reading Tests; Second Language Instruction; Statistical Analysis; Test Construction; Testing; *Test Items; *Test Reliability; *Test Validity; Writing (Composition); *Writing Evaluation
IDENTIFIERS Japan; Test of English as a Foreign Language; Test of English for International Communication

ABSTRACT

This paper traces developments in educational psychology and measurement that led to the Test of English as a Foreign Language (TOEFL) and the test of English for International Communication (TOEIC) and the application of educational measurement terms such as validity and reliability to testing. Use of a table of specifications for planning language tests, procedures for obtaining greater inter- and intra-rater reliability on composition tests are noted. An overview is given of recent criticisms of language testing in Japan. Considerations for item writing are examined briefly, and the major item types in language testing are described and discussed, including multiple choice (in reading and listening), matching (reading), a variety of cloze procedures (reading and grammar), scanning (reading), paraphrasing (reading and grammar), information transfer (reading and writing), editing, guided paragraph writing, paragraphs and essays, question and response (listening and speaking), paraphrase (listening), short talks (listening), and using a table (listening). Some simple statistical procedures for determining item difficulty and for discriminating top-scoring and low-scoring students are also offered. Contains 30 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 397 635

A Survey of Issues and Item Writing in Language Testing

Gregory Strong

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Gregory Strong
Isabelle Makine

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Thought Currents in English Literature
Volume LXVIII, December 1995

The English Literary Society of
AOYAMAGAKUIN UNIVERSITY

FL023728

A Survey of Issues and Item Writing in Language Testing

Gregory Strong

Introduction

This paper traces developments that led to the TOEFL and TOEIC and the application of educational measurement terms such as validity and reliability to testing. The use of a table of specifications in planning a language test is discussed as are procedures for obtaining greater inter-rater and intra-rater reliabilities in composition tests by such means as using a holistic marking scale, sample papers, and rater training. An overview of some recent criticisms of language testing in Japan is presented as well as a review of multiple choice items, four types of cloze reading tests, and other examples of potential questions for reading, writing, and listening tests. Some simple statistical procedures for determining the difficulty of a question and discriminating between top scoring and low scoring students are outlined.

I. Language Tests

English language testing has improved steadily with the introduction of new tests, and refinements in testing administration, and in analyses and critiques of tests, and of particular types of questions or test items. Another factor has been the shift in language teaching methodology. The movement has been away from the classical grammar-translation method and the audio-lingual approach emphasizing listening comprehension to one advocating a communicative class-

room methodology. In this approach, teachers try to engage students in classroom activities where they use the language to actually communicate meaningful information instead of engaging in translation, linguistic analysis, or in repetition and drill activities.

Among the first language tests of English as a foreign language were those developed in Britain. Among these were the Certificate of Proficiency in English (CPE) in 1913 and the First Certificate English Test (FCE) in 1939, introduced by Cambridge University. The university had already began developing national exams in Britain in 1858 (Bachman, Davidson, Ryan, Choi, 1995, pp. 2, 3). Early tests like the CPE and the FCE required language students to write compositions, to translate passages, and to take dictations. The tests and their results were disseminated throughout the British Commonwealth and helped establish standards for many educational programs around the world. Since then, the CPE and FCE have been improved considerably. As well, Cambridge tests of ability at other levels have been introduced, the Preliminary English Test (PET), and the Key English Test (KET) at the lower levels, and the Certificate of Advanced English (CAE) at the level of proficiency between the FCE and the CPE.

The other major development in English language testing came from the United States where language testing began later than in Britain. It started in 1930 in response to rapid increases in the number of student immigrants. The first tests were composed of reading passengers with true and false questions, a short composition, a dictation, an oral test, and a 250–300 word composition (Bachman, Ibid., pp. 3, 4).

New ideas in educational psychology, measurement and testing led to the articulation of a rational process of curriculum design (Tyler, 1949), and the creation of a taxonomy of educational objectives by researchers such as Bloom (1956). In turn, psychometric measures and linguistic principles were applied to language testing. Tests em-

ployed multiple choice items and concentrated on specific lexical and structural points, a focus later to become known as "discrete-point" testing. Educational Testing Services (ETS), Princeton, New Jersey which had been established in 1948, created the Test of English as a Foreign Language (TOEFL) in 1963 to help American universities place foreign students into their programs. In 1991, some 741,000 students wrote the test and more than 2,400 universities in Canada and the U.S. used its scores to place students, making it the single most influential language test in the world (Pierce, 1992, p. 665).

Meanwhile, in Japan, requests from the Japanese Ministry of International Trade and Industry in the 1970s led to ETS developing the Test of English for International Communication (TOEIC) for the Japanese market. Subsequently, this test was used by Japanese corporations in assessing the English language abilities of their employees. Although both the TOEIC and the TOEFL were designed by ETS, the TOEFL uses reading passages and situations found in academic discourse, and the TOEIC employs the language and vocabulary of business English and of commonplace situations.

The STEP test, or Eiken is another test that was introduced in 1963. Devised entirely within Japan by the Society of Testing English Proficiency, it is the most commonly taken test in Japan next to the TOEFL. Over 40,000,000 students have taken it over the last 32 years (Bostwick, 1995, p. 58). Unlike TOEFL, or TOEIC which are norm-referenced tests comparing students, the STEP test consists of six levels of achievement and students either pass or fail at the level they elect to take. In this way, the STEP tests, like the Cambridge proficiency tests, are criterion-referenced. Students either pass or fail a level of English proficiency.

II. Content Validity

Central to improvements in testing have been the two concepts of "content validity" and "test reliability." In short, these are consider-

ations of whether or not a test measures what its designers planned, and to what degree the results of a test would be the same if it were administered again to the same group of students.

A university entrance exam has content validity if it is made up of questions related to the activities and teaching materials used in courses at the university. Of course, not all of the aspects of a program can be covered in a single entrance test. However, representative materials and skills should be part of the test. A test with good content validity will more accurately assess students' abilities in relation to their future studies and place them more appropriately than otherwise.

II. (a) Table of Specifications

Exam specifications are published with each of the major tests discussed earlier, and also by many universities in the U.S., Britain, and Europe that have entrance examinations. These specifications

TABLE OF SPECIFICATIONS

TABLE OF SPECIFICATIONS

CONTENTS	OBJECTIVES					Total
		Identifying the Main Idea	Paragraph Cohesion	Using the Sentence Context	Comprehension	30
	Multiple Choice	10				10
	Matching		5			5
	Cloze			10		10
	Open-ended Question				5	5

(Shoham, 1985, p. 30)

are of great use to students preparing for an exam and help reduce the possibility that they will get high scores by accident instead of by adequate preparation.

Even more useful is a Table of Specifications (Hughes, 1989; Shohamy, 1985) indicating which skills are to be included on a test and by which kinds of test items these skills are to be measured. It is of great assistance in planning tests and in writing test items and discussing them.

In the table, the grid relates a skill or type of knowledge to a question. The table illustrates the specifications for a 30-point reading test. The examinee is being tested for skills in finding the main idea in a reading passage, for reconstructing a narrative, for using the sentence context, and for comprehending the key elements in a reading passage. Each of these skills is cross-referenced with the types of questions that will be used to assess it: multiple choice items, matching, cloze, and an open-ended question with a written response. Ideally, there should be variety in both the skills being tested and in the item types on the test in order to assess a broad range of student abilities. The relative weight of each skill should reflect its importance in the language program.

II. (b) Other Types of Validity

The validity of a test may be measured in a simple, non-statistical way after the students have entered the program. At the end of the term, one could compare the students' classroom results with their scores on the tests. One would expect the highest scoring students to do better than the other students in their classes. If this were the case, then the test would have a high level of content validity because it had predictive validity in indicating students' future scores.

The test would have "concurrent" or "criterion validity" if its results were similar to another test measuring the same skills. It would be expected that two different tests of reading comprehension would

BEST COPY AVAILABLE

show a high degree of correlation between students' marks. The high-scoring students on one test should do well on the second test. However, students' results on a reading test would not likely correlate with their score on a writing test, or on a test of another skill.

III. Reliability

Reliability in testing comes from the concern that there will be inconsistencies in test results and that there always will be a margin of error in reporting test scores. There are five basic types of reliability, several of which can be tested statistically.

The first is a test-retest. This is a hypothetical question about the degree of correlation between test scores if students took the same test twice. In every test, there would be differences in scores due to chance, and perhaps due to error in administering the text. The correlation between the two sets of scores is the degree of reliability of the test's administration. Administering the test twice would have many obvious drawbacks, not the least being that students would recall many of the questions from taking the test initially.

A statistical procedure has been developed to determine this kind of reliability. This is the split halves method. Each examinee is given two scores, one for the even numbered questions on the test and one for the odd-numbered questions. The correlation between the scores on these two half-tests, or the split halves of the test, can be calculated with the Spearman-Brown formula.

The second type of reliability is that of parallel forms which is the extent to which any two forms of the same test measure the same skills or traits. The third is the internal consistency of a test, the degree to which test questions are related to one another and measure the same skills or traits. The internal consistency of a test can be measured through first calculating the standard deviation for the test and then using the Kuder Richardson 21 statistical measure of reliability.

III. (a) Inter-Rater and Intra-Rater Reliabilities

The other two kinds of reliability on a test are "inter-rater reliability" and "intra-rater reliability." These two types of reliability refer to the scoring done on subjective tests. These are open-ended questions requiring written answers, usually paragraph or essay questions. Inter-rater reliability is the degree to which two different raters or markers agree on a score for a student paper. Intra-rater reliability is the extent to which one rater or marker scores consistently from one student's paper to another.

In terms of these latter two types of reliability, there is overwhelming evidence that the scoring of writing is very unreliable unless certain procedures are followed. These procedures include (1) setting the scoring criteria in advance, (2) providing sample answers for the markers, (3) training the markers to use the criteria, (4) scoring each paper twice, and a third time if there is too much difference in the scores attributed to the same paper. These procedures are well-established in the field of English composition research (Braddock, Lloyd-Jones, & Schoer, 1963; Cooper, 1977; Diderich, 1974; Myers, 1980).

To demonstrate the unreliability of marking papers unsystematically, none of these procedures were used in an experiment in the MA TESOL program in the Testing and Evaluation Unit at Reading University, England. In this wellknown experiment, twenty-two MA students scored eight papers between 1 and 20 points (Weir, 1993, p. 155).

The table lists the 22 scorers on the lefthand column. At the bottom of the column is the range of scores assigned to each paper and the mean score for each paper. On the righthand column is the mean score given to the papers by each rater and the range of scores each rater gave to the eight papers.

It can be seen from the table that there is a large range of scores assigned to any one paper. Paper 8# was given a low score of 5 points and a top score of 20. The mean score for this paper is 15

	PAPER NUMBERS									
	1	2	3	4	5	6	7	8	mean	range
RATERS										
A	8	12	12	13	15	8	14	16	12	8–16
B	7	11	12	13	14	7	14	15	12	7–15
C	5	12	11	9	9	4	11	9	9	4–12
D	9	10	14	14	14	6	16	19	13	6–19
E	9	15	15	11	14	8	16	16	13	8–16
F	7	10	11	12	13	14	15	12	12	7–15
G	4	10	15	5	12	3	18	19	11	4–19
H	7	11	10	8	12	6	17	11	11	6–17
I	12	14	17	10	19	10	17	17	15	10–19
J	5	2	3	2	5	1	18	5	5	1–18
K	8	12	14	5	10	13	6	10	11	6–15
L	8	9	11	11	13	9	15	15	11	8–15
M	5	12	15	8	15	9	16	14	12	5–16
N	4	10	12	12	15	3	18	20	12	4–20
O	7	10	10	10	12	15	16	18	12	7–18
P	4	7	12	9	10	3	14	17	10	4–17
Q	5	7	10	8	9	3	11	13	8	3–13
R	3	8	9	9	7	4	17	15	9	3–17
S	8	10	15	10	12	8	15	15	12	8–15
T	3	3	5	5	6	2	8	14	5	2–10
U	12	14	16	13	12	3	19	18	13	3–18
V	10	14	17	14	13	8	18	18	14	8–18
r.	**3–12**	**2–15**	**3–17**	**2–15**	**5–19**	**1–15**	**6–18**	**5–20**		
m.	**7**	**11**	**12**	**10**	**12**	**7**	**15**	**15**		

(Weir, 1993, p. 155).

points suggesting that it is a good, passing paper because so many raters gave it a high score. However, rater J scoring it at 5 points fails it. Even the smallest range of marks for a paper is considerable. Paper 1# has a mean score of 7 points and is likely a poorly written paper. It was given a low score of 3 and a top score of 12 which not

only passes the paper, but is a higher score than the score some raters gave Paper 8#. It can be seen that there is little inter-rater reliability between the markers.

This experiment indicates that even well-educated, experienced markers with expertise in EFL such as these graduate students will score papers inaccurately without adequate criteria and rating procedures.

Although this example does not demonstrate the problem of intra-rater reliability from one paper to another, this has been well-established in the research literature in composition in a first language. Coffman and Kurfman (1968) show that marking behaviour in a single rater changes over the marking period. This also is well-established by others (Braddock, Lloyd-Jones, & Schoer, 1963; Cooper, 1977; Diderich, 1974; Myers, 1980).

III. (a) (i) A Holistic Scale

These researchers (Ibid.) suggest the use of a holistic or general impression marking scale for scoring papers. The markers form a "holistic" or general impression of each paper's content, organization, sentence structure or style, and its written expression or use of grammar. The scales used are commonly five-point, six-point, or twelve-point scales. The smaller the range of scores on a scale, the greater the reliability in marking. This is because it is more likely that two raters will assign the same score to a paper if they are using a five-point scale than a twelve-point one. Afterward, the students' marks for that portion of the test can be scaled to represent a larger portion of their exam marks than five or twelve points.

One of the better known scales currently in use is the one developed by ETS for use with the Test of Written English (TWE). The TWE was developed to meet the need for an essay test in some university admission requirements.

This six-point scale was modified by Strong (1990) and subse-

quently employed by the English Department of Aoyama University in the writing assessment portion of the placement test of the Integrated English program in 1995. There are six bands on the scale. There is a description of the content, organizational patterns, the use of paragraph transitions, and effective sentence structure, and grammar for each band. The bands are as follows: (6) Advanced student writer, (5) Good student writer, (4) Competent student writer, (3) Modest student writer, (2) Marginal student writer, (1) Limited student writer with descriptors for each band that outline the general features of a paper at that level.

To properly train markers in using the scale, an outline of a complete answer at band 6 is devised. Then a committee selects a series of papers randomly and chooses among them for six anchor papers that the committee feels demonstrate the writing competencies at each of the different bands on the scale. Afterward, the raters examine the six anchor papers and try to determine where each fits on the scale. The raters discuss their reasons for assigning their marks, and then they compare their results with those of the committee.

Raters are asked to mark on general impressions and to avoid deducting points for individual grammatical errors such as spelling mistakes, or instances of incorrect subject-verb agreement, or any lack of topic sentences. The raters are to ask themselves if a paper that may be written by an Advanced student writer is thoughtful, well-organized, and has only minor errors, or if the paper seems to be written by a less advanced writer and fits elsewhere on the scale.

A head rater works with small groups of raters, randomly checking each rater's marked papers to determine if the rater has been using the scale correctly. Each paper is marked twice. If there is more than one point difference between the scores on a paper, then it is scored by a third rater, and usually the three scores are averaged. Once teachers are trained in using the scale, marking proceeds quickly and accurately with only a few minutes spent on each paper. There are no

6

Advanced student writer
-logical and persuasive argument
-well-organized paragraphs
-thoughtful ideas, names, details
-appropriate transition words
-minor errors in grammar and punctuation
-interesting word choice

5

Good student writer
-argument is clear although obvious
-an organized paragraph
-suitable examples
-few transitions and less varied sentences
-errors in grammar and punctuation don't interfere with communication

4

Competent student writer
-an argument is apparent
-one or two developed examples
-simple transitions
-grammatical errors sometime interfere with communication

3

Modest student writer
-badly organized paragraph
-underdeveloped examples
-repetitive word choice
-minor and major errors in grammar
-repetitive sentence structure

2

Marginal student writer
-question answered very superficially
-at times seems incoherent
-underdeveloped paragraph
-flawed sentence structure
-very limited word choice

1

Limited student writer
-inability to comprehend the question
-severely underdeveloped paragraph
-obscured meaning in the sentences
-persistent major grammatical errors

BEST COPY AVAILABLE

comments or corrections made on any of the papers.

IV. Toward Improvements in Validity and Reliability

Aside from the statistical analyses to determine test validity, and the treasures to improve reliability suggested earlier, there ate a number of steps that can be taken to ensure better testing. Hughes (1989) outlines these:

1. Plan the test systematically.
2. Include a variety of item types on a test to assess a broad range of language skills.
3. Identify candidates by number, not name.
4. Do not allow candidates choices of questions as this makes it harder to compare candidates.
5. Write test items with clear expectations.
6. Provide good instructions, possibly in the candidates' native language.
7. Ensure that tests are well laid out and completely legible.
8. Familiarize candidates with the format and testing techniques in advance, and provide sample questions.
9. Provide testing conditions that are uniform and not distracting to the participants.
10. Use items that encourage unambiguous scoring where possible.
11. Provide a detailed scoring key specifying acceptable answers, and noting the points to be assigned for partially correct answers.
12. Train raters where the scoring is subjective.
13. In the case of subjective items such as open-ended questions, and extended writing, agree on the appropriate answers and scores before marking the tests. Use sample papers and training sessions for these questions.
14. Where testing is subjective, especially in paragraph and essay tests, use two raters.

(pp. 36–42.)

In describing the benefits of language testing, Brown (1995) notes that tests can be used to sort students according to their language abilities and create more homogenous classes which will be easier to teach. Brown maintains the tests should be adapted from existing tests, or developed exclusively for an institution in order to select the students most suitable for its programs.

In this area, Japanese colleges and universities deserve considerable recognition for developing tests that are unique to each institution. Furthermore, the tests themselves are created cooperatively in exam committees and there is discussion and criticism of test items. These features of language testing in Japan are very positive ones.

However, Brown (1995) and other researchers (Bostwick, 1995; Brown and Yamashita, 1995) have several criticisms of examinations in Japan. Brown and Yamashita (Ibid.) analyzed the entrance exams at 21 private Japanese universities including Aoyama, Keio, Rikkyo, Sophia, and Waseda, and 10 public universities, among them, Kyoto, Osaka and Tokyo universities. The sources for their study were two commercially available books, Koko-Eigo Kenkyu (1993), *'93 Shiritsu Daigaku-ben: Eigo Mondai no Tettiteki Kenkyu*, Tokyo, Kenkyusha and Koko-Eigo Kenkyu, *'93 Kokukoritsu Daigaku-ben: Eigomondai no Tetteiteki Kenkyu*, Tokyo, Kenkyusha.

Brown and Yamashita (1995) based their analysis on exam item types, and the comparative difficulty of reading passages on exams. They used a computer spreadsheet program to code and count the types of questions on the different university exams. Afterward, they used the Que computer software (1990) *Right Writer: Intelligent Grammar Checker* (version 4.0), Sarasota, Florida to analyze features in the reading passages on the exams. This software program calculates the number of words, the syllables per word, the number of words per sentence, and the number of sentences. It also determines the readability of passages using the Flesch, Flesch-Kincaid, and Fog readability indexes.

Among the observations they made were that there were substantial differences between the reading sections of the exams. The public universities tended to have more reading passages, but of shorter length. The reading difficulty of the passages ranged from those appropriate for native speakers at sixth grade, in the case of Kansai University, to those suitable for third year university students in the case of the entrance exam at Nagoya University (Ibid., p. 89). As for item variety, Kangai and Sophia universities placed a heavy emphasis on multiple choice items while other universities such as Kyoto emphasized translation (Ibid., p. 91). Furthermore, only four universities, Aoyama University and Tokyo University among them, included listening items on their exams. This was despite recent Monbusho guidelines advocating more listening and speaking activities in English instruction in Japanese junior and senior high schools (Ibid.).

The researchers made several additional observations. New sets of directions had to be given often in exams. Test lengths also varied considerably as well. They suggested that students taking these exams would be confronted by too much variation in language testing and that this situation might discriminate in favour of students who were more test-wise rather than those who were better at using English. The researchers also suggested that translation activities, besides being hard to grade, might be too difficult a skill to require of students with only limited English study in junior and senior high school.

Finally, Brown and Yamashita (1995) criticize the universities in their study because none of them do any of the statistical analyses of reliability and validity of their language tests that are common practice elsewhere. They suggest that Japanese universities either follow the guidelines established by the Committee to Develop Standards for Educational and Psychological Testing. (1985). *Standards for Educational and Psychological Testing,* Washington, D.C.: American Psychological Association or adapt these to Japan (Brown & Yamashita, Ibid., p. 98).

Bostwick (1995) makes a similar criticism of the Eiken STEP and the Jido Eiken STEP tests. He argues that although they are proficiency tests, there is no information available on their validity and reliability. There is no explanation of how levels or passing scores are calculated. As a result, it is not possible to learn whether the tests successfully distinguish between several levels of language performance and whether these levels are consistent from test administration one year to the next.

V. Item Writing: Reading, Writing, Listening

Under the impact of the communicative language teaching methodology, language test items are changing. Test items in the past were almost exclusively of the discrete-point type where specific language points such as vocabulary items, and verb conjugations were tested. But now tests include language tasks where students complete activities that include several different language skills and may be based on real-life activities such as reading signs, and brochures, following directions, note-taking, and writing different kinds of compositions such as paraphrases, summaries, and statements of opinion (Shohamy, 1985).

In addition, many test items used to be based on indirect measures of language ability such as a knowledge of grammar being used to test a student's writing ability. These items are being replaced by more direct measures of writing such as requiring students to write compositions.

Productive language skills such as speaking and writing also are being tested more extensively than before. Both skills are measured in such contemporary tests as the Cambridge series of language proficiency tests mentioned earlier, and the CANtest, a Canadian-developed test of language skills created at the University of Ottawa. The same is true of the TOEFL which has introduced two additional tests, the Test of Written English (TWE), and the Test of Speaking

English (TSE). Furthermore, ETS is planning a major revision of TOEFL exams in the TOEFL Year 2000 project to change the examination into a more communicative, task-oriented one (Brown & Yamashita, 1995).

In general, several considerations apply when designing test items in reading and listening. According to Shohamy (1985) and Hughes (1989), these comprise (1) the importance of including different types of reading texts, or listening materials, (2) the use of authentic texts, and of real-life tasks wherever possible. Finally, (3) item designers should not attempt to find too many questions about a single reading passage or listening text. The use of a broad range of subjects and types of questions provides each examinee with what Hughes (Ibid.) calls "fresh starts" and taps different language abilities.

The remainder of this paper will outline some of the major types of items in language tests. These are items used in assessing reading, writing, and listening skills and do not include oral interviews and tests.

V. (a) Multiple Choice (Reading and Listening)

The continued attraction of multiple choice test items lies in their unambiguous answers, the comparative ease with which they are scored, and their statistical reliability. They are now used for a variety of question types of reading and listening skills. But their original purpose was for assessing terminology, facts, classifications, and other discrete areas of knowledge (Gronlund, 1977).

A multiple choice item consists of a stem, a correct answer, and three or four alternatives or distractors. As far as possible, the stem should be written in simple, clear language and most of the wording of the question should be in the stem. The item difficulty is controlled by varying the problem in the stem or by changing the alternatives. The answer and the distractors must be grammatically consistent with the stem and parallel in length and grammatical structure

and the distractors must all be plausible answers to the uninformed (Ibid., p. 45, 49).

The problem with multiple choice items is that they encourage guessing and a student could score as high as 33% just by chance (Hughes, 1989). Although guessing is a factor on other test items, the effect is much less. Hughes (Ibid.) also notes that the item restricts what may be tested and that it is very difficult to write plausible alternatives to the correct answer.

However, Heaton (1988) contends that these test items can still be effective in discriminating between students, especially if they are pre-tested on a representative sample of the test population. The latter precaution will help in gauging the difficulty of the test and can be used to compare a test with those of previous years. Heaton (Ibid.) counters the criticism of guessing by the observation that examinees rarely make wild guesses, but usually base their choices on partial knowledge of a question anyway (p. 28). Heaton recommends four distractors for grammar questions, and five for vocabulary and reading questions (Ibid.).

V. (b) Matching (Reading)

Typically, the matching question is a modification of the multiple choice form where all the stems or premises are listed in one column on the right and a longer list of distractors, called responses is listed in a column to the right (Gronlund, 1977). In matching questions, the lists should be short, and each response should be a plausible alternative for all of the premises. The factor of guessing is reduced in this type of question because there are so many possible answers.

One of the better known applications of matching test items is as a test of vocabulary and the use of context clues. Given a list of words at the end of a passage, students are asked to find synonyms in the passage itself. A detailed context is supplied by the passage making this an economical method of testing vocabulary.

group *band*
owned *exclusive*
specific *particular*

THE TEHUELCHES
The Tehuelches lived in a band -- usually of between fifty and a hundred people. Each band had exclusive rights to a particular hunting area...

(Heaton, 1988, p. 60)

Another application for matching questions is in a test of reading comprehension. The examinees have to select the appropriate phrase in order to create a cohesive expository passage. They are given several sentences at the beginning of the passage and at the end as well.

IN SEARCH OF LANGUAGE'S MISSING LINK

American linguists believe they are approaching their profession's ultimate goal -- the reconstruction of the 'mother tongue', the language spoken by earliest humanity. The ancient words of the first human beings are about to be heard again, they say...

...The human race was at that time just a loose band of people inhabiting a region of sub-Saharan Africa. __1__, replacing neanderthals and other rivals, bearing our language round the world.

As humanity spread out, this mother tongue divided into various dialects which in turn developed into new languages. __2__, leading to the development of modern mankind's many different languages ranging from Aborigine to Eskimo, from Serbo-Croat to Basque...

A. Then we emerged, out of Africa, to take over the world
B. This process was repeated over the centuries

(Cambridge, 1991, CAE Paper 3, p. 6)

In a similar type of question item, students may be asked to find the appropriate sentences to create a cohesive narrative text. Both types of test items require that the difficulty of the reading passage be appropriate for the students being tested and that the responses be thoroughly pre-tested, preferably on a sample group of students.

BEST COPY AVAILABLE

RARITY

As we threaded our way down and down, one of our party stopped and knelt. There, hanging downwards on a dead branch on the forest floor, was what looked like a large, dried and blackened flower, two petals partially open. I was about to move on when the petals quivered weakly and a bright, unwinking eye gazed at me from the flower's base.

1	

I gently cupped my hand around the swift and lifted it, wet and shivering minutely. Obviously it had been knocked out of the sky by the recent storm. Falling drenched and helpless into the forest, the bird had tried to regain its habitat by climbing the branch -- a brave but hopeless attempt...

A

In that instant the argument between the scientist and the conservationist in me was decided.

B

Abruptly the image reversed itself, as illusions do, and it became a bird...one of the swifts -- incongruous, the most aerial of all birds, stranded deep in the forest. It was as strange as finding a whale. How had it reached this nadir?

(Cambridge, 1991, CAE, Paper 1, p. 4, 5)

V. (c) Cloze Tests (Reading and Grammar)

In cloze tests, the examinees are given a passage from which words have been replaced by blanks and they have to decide which word best fits each blank. The more skilled the language learner, the better able he or she will be at choosing the best word for each blank. When a reading appropriate to the level of the students is chosen, this test has a high degree of reliability.

One of the advantages of cloze tests is that an open-ended cloze test (as opposed to a multiple choice cloze test with distractors) is an easy test to construct. It can also be used as a effective substitute for grammar tests because students are given an actual language sample and are presented with a full range of structural questions from verb choice, and use of tense, prepositions, and articles to questions of

semantics and rhetoric.

Cloze passages usually are constructed by leaving the first few sentences of a reading intact. Several sentences are left at the end of the passage in order to provide a complete context.

Often, there is a total of 30–50 blanks left to be completed (Ikeguchi, 1995, p. 168). And cloze passages are of four types: fixed rate, rational deletion, multiple choice, and c-test.

V. (c) i Fixed Rate Cloze

In this kind of cloze test, the test items are created by deleting words at regular intervals, every fifth word, or more commonly, every seventh word. The more frequently words are deleted, the more difficult the passage becomes (Brown, 1988, 1983).

V. (c) ii Rational-Deletion Cloze

In this type of cloze test, different types of words are deleted to test different aspects of the examinee's knowledge of English. To find out the answers to the test items, candidates must look within the clauses where the blank appears, within the sentence, or within the paragraph itself. In this manner, this test is of students' abilities to read at semantic, syntactic, and paragraph levels of comprehension.

ECOLOGY
Water, soil and the earth's green mantle of plants make up the world that supports the animal life of the earth. Although modern man seldom remembers the fact, he could not exist without the plants that harness the sun's energy and manufacture the basic foodstuffs he depends ___*upon*___ for life. Our attitude ___*toward*___ plants is a singularly narrow ___3___. If we see any immediate utility in ___4___ plant we foster it...

(Hughes, 1989, p. 66)

V. (c) iii Multiple Choice Cloze

In construction, this type of cloze test is either the fixed rate deletion or rational deletion type. It has the same advantages and

BEST COPY AVAILABLE

disadvantages of multiple choice items. Because the choice of potential answers is supplied, students finish the test far more quickly than if they were doing an open-ended cloze test.

But the multiple choice cloze test is far more difficult to create than an open-ended test because the distractors must be written as well. This usually requires pre-testing to select suitable distractors. A quick, effective way to create these distractors is to give the test as an open-ended cloze test to a sample population and use their responses as the basis for test items. Alternately, one might write distractors that all use the same part of speech as the correct answer. Either method has been found to have a high degree of reliability (Ikeguchi, 1995).

V. (c) iv Modified C-Test

As with the other cloze tests, the first few sentences and the last few sentences are left intact in the passage to give the examinees a complete context. The c-test is a grammatically-based, modified cloze test where the second half of every second word is deleted, (excluding numbers and proper names). The c-test is very easy to construct, and although open-ended, is easy to score because there is usually only one acceptable answer for each question.

A FIRE ENGINE CREW
There are usually five men in the crew of a fire engine. One *of*___ them dri*ves*__ the engi*ne*__. T*he*__ leader h_5_ usually be_6_ in t_7_ Fire Ser_8_ for ma_9_ years. H_10_ will kn_11_ how t_12_ fight diff_13_ sorts o_14_ fires. S_15_, when t_16_ firemen arr_17_ at a fire, it is always the leader who decides how to fight a fire. He tells each fireman what to do.

(Klein-Braley and Raatz, 1984)

To improve on its reliability over other cloze tests, a c-test usually includes about six different short passages in a test with about 100 deletions altogether (Ikeguchi, 1995). Narrative and explanatory tests tend to be more accurate than passages of argument and description

and the reading level should be appropriate for the students being tested. Adult native speakers should obtain virtually perfect scores on a good c-test, however, language students will find the same test very, very difficult to do. Their scores are often as low as 50% of the questions correctly answered (Ikeguchi, 1995).

V. (d) Scanning (Reading)

Scanning questions assess a subskill of reading, that of scanning for information. In scanning, students are trying to locate factual information rather than to comprehend a text. Some questions should be inferential as well such as calculating the age of someone or determining the emotional tone of a piece of writing. A scanning test must be timed and students allowed only a fixed period of time in order to find the answers. Generally, a minute per question on a test of ten items is allowed. Once the time has elapsed, the papers are taken away.

Almost any kind of authentic reading material provided for native speakers of English is suitable for this kind of test including encyclopedia references, newspaper articles, brochures, advertisements, letters, and notices (Shohamy, 1985). The reading passage and questions should be pre-tested with a sample group in order to establish the difficulty of the passage. Because the questions are open-ended, an answer key must be prepared with the correct answers and any potential alternates.

THOMAS GRAY

Thomas Gray who was born in London in 1716. and died in Cambridge in 1771 was the poet who wrote "Elegy in a Country Churchyard". It is one of the best known of English poems. Although Gray wrote very little, he was a dominant figure in the mid-18th century. His work initiates the Romantic period.

Born into a prosperous, but unhappy family, Gray was the only survivor of twelve children. His father was harsh and violent and his mother was long-suffering...

1. **How old was Gray when he died?**
2. **How many children died in his family?**

V. (e) Paraphrase (Reading, and Grammar)

This item type requires students to write a sentence that is equivalent in meaning to the example. Part of the paraphrase is given in order to restrict the students to a particular grammatical structure. This kind of grammatical test item is an effective alternative to testing students' knowledge of grammar with multiple choice items. It requires them to actually produce particular grammatical structures. An answer key would have to be developed for this question and a decision made over partial marks to be awarded for certain answers.

1. Testing the passive, past continuous form.
When we arrived, a policeman was questioning the bank clerk.
When we arrived, the bank clerk .....

2. Testing the present perfect with *for*.
It is six year since I last saw him.
I.....six years.

(Hughes, 1989, p. 143)

V. (f) Information Transfer (Reading and Writing)

This type of item combines reading and writing skills because examinees are given information in the form of an office memo and then required to complete a summary. There may be several possible answers, therefore an answer key should be constructed and tested before this item is scored.

THE JULY TRAINING COURSE

I've just phoned the Personnel Department and got details of that Training Course we were interested in.

It's going to last a couple of days and will cover a whole lot of things. They've decided that on the first morning they'll have someone talking about what's new in the profession.

After lunch people will be discussing things in groups and then they'll show the training video.

They're very sorry but Angela Gresty can't do the opening speech and they're having a speaker from the Chamber of Commerce instead. Anyone in the office can go along on either or both days.

Do you think it sounds useful?

JULY TRAINING COURSE

The (two-day) course will cover a (2)_____ of (3)_____. It (4)_____ with a (5)_____ new (6)_____ in the profession. After lunch there will be group (7)_____, (8)_____ by the training video.

(9)_____, Miss Gresley is (10)_____, but (11)_____ will be a speaker from the Chamber of Commerce instead. Office (12)_____ are invited to (13)_____ the course on either or both days.

(Cambridge, 1991, CAE, Paper 3, p. 5)

V. (g) Editing

As with scanning, this kind of test item assesses a subskill, in this case the writing subskill of editing. It is not actually a productive activity such as writing a composition. But this test does measure an

	THE NEED FOR SLEEP
0	**Although sleep isn't yet fully understood, the massive**
0	**amount of research that has been done in the last ten**
1	**years makes it look as if the rest it gives in the mind is**
2	**probably more important than the rest it gives the body.**
3	**The amount of sleep one needs varies from person to**
4	**person as well as from time to another time. Extroverts**
5	**sleep less and introverts sleep more than average. You**
6	**probably need to more sleep at times of stress, when**
7	**you're dieting or when you want change jobs. You need**
8	**less sleep when your life is running smoothly. While a**
9	**recent study shows that short sleepers are more likely to**
10	**be efficient, energetic and ambitious about their lives;**
11	**they deal with no worries by keeping busy. Long**
12	**sleepers (nine hours plus) tend neither to be very original**
13	**and critical. They may be artistic and creative but they**
14	**may not be very sure of themselves, their career in**
15	**choices or their lifestyles. Sleeping problems are**
16	**particularly common in the middle age...**

(Cambridge, 1991, CAE, Paper 3, p. 5)

aspect of writing ability. It is easily created by taking a non-fiction passage and adding unnecessary words. Students' answers are scored against an answer key.

V. (h) Guided Paragraph Writing

In this type of writing test, students are given an outline for their work. This makes this item a test of their ability to use notes to generate ideas, and to develop their ideas into written form. As with other writing tests, a model paper should be devised, and a marking scheme [illegible] for the answer.

Write a descriptive paragraph of about 75 words about a store or business that you know very well. Base your paragraph on answers to the following questions:

1. What is it called?
2. When did it start to do business?
3. How many employees does it have?
4. What do the employees have to do?
5. Does it have a lot of customers/clients? Why(not)?
6. Why do you choose to go there rather than somewhere else?
7. Is it a good example of what such a store/business should be?

(Madsen, 1983, p. 111)

V. (i) Paragraphs and Essays

As Shohamy (1985) notes, open-ended written questions requiring a written response are among the easiest test items to construct. However, as described earlier in this paper, there are several procedures that teachers marking the exam should take such as devising a marking scale, and training raters how to use it with sample papers.

Paragraphs and essay questions could include describing something, comparing and contrasting ideas or objects, and expressing an opinion about a problem. Diagrams or labelled pictures also could be supplied to give students the vocabulary to make a comparison between two objects, such as a bicycle and a motorcycle. Ideally, writ-

ing topics should be about things that are familiar to all the examinees in order to ensure that every student has an equal chance to demonstrate their writing ability.

1. Expository Writing
What are the qualities of an ideal parent?

2. Comparison/Contrast
Compare and contrast student life in high school and university.

3. Expressing an Opinion
Hiro and his two friends, Manabu and Toru from university were out drinking Friday night. Hiro was driving his mother's car and they had an accident. Luckily, no one was hurt, but repairs to the car will be very expensive.
Do you think Hiro should pay for all the repairs himself or should his friends pay, too?

V. (j) Question and Response (Listening and Speaking)

This type of item usually appears on the TOEIC. It is primarily a test of listening, but it does require some knowledge of spoken English as well. Therefore, it is a good addition to a listening test. However, past versions on the TOEIC have included only three distractors for each sentence, therefore guessing has been encouraged.

QUESTION AND RESPONSE

Q: Have you been working for IBM long? (tapescript)
a) About 8 years now.
b) I began working a long tome ago.
c) Until 62.

(Pifer, 1981, p. 4)

V. (j) Short Conversations (Listening)

In this type of item which also appears on the TOEIC and in similar forms on many other listening tests, the examinee has to determine the location, subject, and speaker in a short conversation. There is only one question for each conversation. The conversations are drawn from everyday activities and there are about 30 different

BEST COPY AVAILABLE

questions on the TOEIC. On the TOEFL, the conversations are longer and more questions are asked about each conversation. The same is true of the listening sections of other tests.

SHORT CONVERSATION
(Tapescript)
W: Let's see, you have two pairs of slacks and one sports jacket. How soon do you need them?

M: By Friday. Also, there are a couple of grease spots on this pair of pants. Can you take them out?

W: Sure, that's no problem. Okay, you can pick them up any time after Friday...

Where does this conversation take place?
a) In a department store.
b) In a supermarket.
c) In a dry cleaning shop.
d) In a tailor shop.

(Pifer, 1981, p. 3)

V. (k) Paraphrase (Listening)

In this item, students choose the correct paraphrase of the dialogue. To ensure standard answers, multiple choice questions are used.

PARAPHRASE
(Tapescript)
M: I ate too much dinner tonight.
W: You told me you were going on a diet.
M: I know, I know, but I couldn't stop eating.

Q: What did the woman expect?
a) She expected to lose weight.
b) She talked about her new diet.
c) She could not stop eating.
d) She expected her friend to diet.

(Pifer, 1984, p. 7)

V. (l) Short Talks (Listening)

These items are monologues, recorded speeches, news and radio

broadcasts, and public announcements. They are meant to be as authentic as possible and once again form part of the listening section of many different tests. Following each one are multiple choice questions, open-ended questions or information tables. The examinees hear the tape once or twice, depending on the group and on the difficulty of the item. Then they read and answer the multiple choice questions.

NEWSCAST
(Tapescript)
Another terrorist attack was reported this morning when an explosive device went off in a crowded commuter train. Six people were killed and 62 others injured. This is the 4th bomb blast since the beginning of the New Year. Police officers have said that they will step up security precautions by inspecting all trains and passengers.

1. What is the man reporting?
a) A plane accident.
b) A gun battle.
c) An attack on innocent civilians.
d) A train crash.

2. What caused the death and injuries?
a) A bomb.
b) Pilot error.
c) Panicked crowds.
d) A collision.

(Pifer, 1981, p. 16)

V. (m) Using a Table (Listening)

In this kind of listening test, there is an information transfer where candidates listen to monologue, a lecture, directions, or a conversation and record the information on a table. The questions are open-ended and there is none of the guessing associated with multiple choice questions. An answer key is developed for the table and decisions made over partial marks to be awarded.

In the following example, examinees listen to a tape about a student's schedule and fill out a calendar. They start with the first statement about Monday, the 12th, and then work forward and then back-

BEST COPY AVAILABLE

ward listening to such expressions such as "today," or "last week."

A BUSY MONTH'S SCHEDULE
(Tapescript)
1. Today is Monday, the 12th. It's a very busy day. I have two classes, English, and French, and my club meeting.

2. Last week, our class has a party on Friday...

M	T	W	T	F	S	S
		1	2	3	4	5
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

(Corchene, 1991)

VI. Conclusion

The research on language testing indicates that test development is very much an on-going process of item design and refinement. As the methodology of language teaching changes, so do language tests. At the same time, new procedures for analyzing readability and calculating statistics for item difficulty and item discrimination between students can improve language testing.

In the full process of test development described by Brown (1995), each question should be examined after the test, or preferably after pre-testing with a sample population. First, the test questions should be analyzed to see whether they discriminate well between high ability and low ability students. The facility index is calculated by

adding the number of students who correctly answered a particular test question and dividing that by the number of students who undertook the test. If 40 of 200 students answered a question correctly, then the facility index for the question would be 40/200 = .20 indicating a fairly difficult question as only 20 percent of the students got it correct (Ibid., p. 43). According to Brown (Ibid.), an ideal question for a norm-referenced test is one that 50–60 percent of the students answer correctly and that anything outside facility values between .30 and .70 is either too hard or too easy.

Secondly, another statistic should be calculated for each question, that of the discrimination index for each question. This examines the degree to which high ability, or top scoring students answered a question correctly compared to low scoring students (Ibid.). To calculate this statistic, student scores for both the test and the question must be tabulated. The students are ranked in terms of a top third of the students taking the test and the bottom third. Then the percentage of students in the top scoring group who got the question right, for example, 80 percent, is subtracted from the percentage of students in the bottom scoring group who got the question right, perhaps 20 percent. The answer would be an acceptable 60 percent discrimination value between the high scoring and low scoring groups.

New technology in the from of computer software can assist test designers in determining readability and item difficulty and discrimination. Word Perfect 6.0 and MS Word 6.0 are among the many word processing programs that offer the Flesch-Kincaid and Fog analyses of readability. Potential test passages could be scanned using a computer, saved as files, and then analyzed early in the test design process. As for determining item difficulty and discrimination, the new LXR · Test 5.1 which runs on both IBM and MacIntosh platforms will perform both these functions in addition to scrambling questions, and renumbering them for future use. The program gener-

ates different test versions, can import question data, graphics, and can produce machine markable tests. This particular software package, developed in California, is highly rated and is widely used in high schools and colleges in western Canada and the western United States (University of Calgary, 1994)

The literature on language testing indicates that once a test has been designed, and the items written, it should be thoroughly critiqued. Not only should members of the test committee take the test before it is administered, but other colleagues as well. Ideally, the test should be pre-tested with a sample group. But if this is not possible, then after the test is administered, it should be thoroughly analyzed using statistics such as the facility and discrimination indexes described earlier. Good items can be banked for a part of some future test. An even more compelling reason for test analyses is that much can be learned about the appropriateness of test items for a particular group and that good language tests set the standards for subsequent ones.

Ultimately, good language testing is an essential tool in creating effective language teaching programs. The more valid and reliable the test, the better it will be in assessing students' abilities and in placing them accurately into a program or in assessing their work.

VII. Bibliography

Bachman, L. Davidson, F. Ryan, K. Choi, I. (1995). *An Investigation into the Comparability of Two Test of English as a Foreign Language: The Cambridge-TOEFL Comparability Study.* Cambridge: University of Cambridge.

Bloom, B. (Ed.) (1956). *Taxonomy of Educational Objectives: Cognitive Domain.* New York: David McKay.

Bostwick, R. (1995). Evaluating Young EFL Learners: Problems and Solutions. In J. Brown, S. Yamashita (Eds.) *Language Testing in Japan.* Tokyo: The Japan Association of Language Teachers, 57–63.

Braddock, R., Lloyd-Jones, R. & Schoer, L. (1963). *Research in Written Composition.* Urbana, IL: National Council of Teachers of English.

Brown, J. (1995). Developing Norm-Referenced Tests for Program Level Decision Making. In J. Brown & S. Yamashita (Eds.) *Language Testing in Japan.* Tokyo: The Japan Association of Language Teachers, 40–47.

BEST COPY AVAILABLE

Brown, J. (1985). Cloze Procedure: A Tool for Teaching Second Language Reading. TESOL Newsletter 20, 5, 1 & 7.

Brown, J. (1983). A Closer Look at Cloze: Validity and Reliability. In J. Oller, Jr. (Ed.) *Issues in Language Testing Research.* Rowley, MA: Newbury House, 237–250.

Brown, J. Yamashita, & S. (1995). English Language Entrance Exams at Japanese Universities: 1993 and 1994. In J. Brown & S. Yamashita (Eds.) *Language Testing in Japan.* Tokyo: The Japan Association of Language Teachers, 86–100.

Cambridge Examinations, Certificates, and Diplomas. (1991). (CAE) Certificate in Advanced English. Cambridge, U.K: University of Cambridge Local Examinations Syndicate.

Cooper, C. R. (1977). Holistic Evaluation of Writing. In C.R. Cooper, & L. Odell (Eds.) *Evaluating writing: Describing, Measuring, Judging.* Urbana, IL: National Council of Teachers of English.

Corchene, R. (1991). [CANtest Listening Practice]. A Calendar. (Cassette Recording), Beijing, Canada-China Language Project, Beijing Normal University.

Clankie, S. (1995). An Introduction to Commercial English Tests in Japan. *The Language Teacher,* 19, *4,* 8–10.

Coffman, W., & D. Kurfman. (1968). A Comparison of Two Methods of Reading Essay Examinations. *American Educational Research Journal,* 5, *1,* 11–120.

Diederich, P. (1974). *Measuring growth in English.* Urbana, IL: National Council of Teachers of English.

Gronlund, N. (1977). (2nd Ed.). *Constructing Achievement Tests.* New Jersey: Prentice Hall.

Heaton, J. (1988). *Writing English Language Tests.* Hong Kong: Longman.

Hughes, A. (1989). *Testing for Language Teachers.* New York: Cambridge University Press.

Ikeguchi, C. (1995). Cloze Testing Options for the Classroom. In J. Brown & S. Yamashita (Eds.) *Language Testing in Japan.* Tokyo: The Japan Association of Language Teachers, 166–178.

Klein-Braley, C. & Raatz, U. (1984). A Survey of Research on the C-test. *Language Testing,* 2, 76–104.

Madsen, H. (1983). *Techniques in Testing.* New York: Oxford University Press.

Myers, Miles. (1980). *A Procedure for Writing Assessment and Holistic Scoring.* Urbana, IL: National Council of Teachers of English: ERIC.

Pierce, B. (1992). Demystifying the TOEFL. *TESOL Quarterly,* 26, 4, 665–689.

Pifer, G. (1981). *Practice Tests for TOEIC.* Surrey, England: Nelson.

Shohamy, E. (1985). *A Practical Handbook for Language Testing for the Second Language Teacher.* Ramat Aviv: Israel: Shaked.

Strong, G. (1990). A Comparison Group Study on the Effects of Instruction in Writing Heuristics on the Expository Writing of ESL Students. (Unpublished master's thesis), University of British Columbia.

Tyler, R. (1949). *Basic Principle of Curriculum Design.* Chicago: University of Chicago.

University of Calgary. (1994). Testing Software: A Review. (LXR Test 5.0) *New Currents in Teaching and Technology,* Information Services and the Teaching Development Office: University of Calgary, 1, 1, November.

* The LXR Test 5.0 (Logic eXtension Resources) computer software program is available from many educational software distributors such as seabyte @ aol. com, seabyte Educational Consultants. 1515 Mortimer st., Victoria, B.C, Canada V8P 3A3.

Weir, C. (1993). *Understanding and Developing Language Tests.* New York: Prentice Hall.

Weir, C. (1990). *Communicative Language Testing.* New York: Prentice Hall.